Mining Significant Words from Customer Opinions Written in Different Natural Languages

Jan Žižka and František Dařena

Department of Informatics/SoNet Research Center
Faculty of Business and Economics, Mendel University in Brno
Zemědělská 1, 613 00 Brno, Czech Republic
{zizka,darena}@mendelu.cz

Abstract. Opinions expressed by text documents freely written in various natural languages represent a valuable source of knowledge that is hidden in large datasets. The presented research describes a text mining-method how to discover words that are significant for expressing different opinions (positive and negative). The method applies a simple but unified data pre-processing for all languages, providing the bag-of-words with words represented by their frequencies in the data. Then, the frequencies are used by the algorithm which generates decision trees. The tree decisive nodes contain the words that are significant for expressing the opinions. Positions of these words in the tree represent their significance degree, where the most significant word is in the node. As a result, a list of relevant words can be used for creating a dictionary containing only relevant information. The described method was tested using very large sets of customers' reviews concerning the on-line hotel room booking. For more than 15 languages, there were available several millions of reviews. The resulting dictionaries included only about 200 significant words.

Keywords: textual documents, multilingual documents, natural language, opinion analysis, text mining, decision tree, significant attribute.

1 Introduction

The research goal was to find a method how to create a dictionary that contains only words significant for expressing opinions. One of typical electronic text utilizations is collecting and analyzing expressions of opinions provided by people that used some services or purchased some goods. The more expressions are available, the more valuable information and knowledge can be revealed inside, which can be later used both in commercial and non-commercial areas [9]. On the other hand, very large volumes of data need processing by machines, and, in addition, data having the form of unstructured natural language documents are generally difficult for such processing [1]. Further, the problem is intensified when customers use many different languages because there is not an unified method developed for the easy processing of such data. Particularly the commercial branch is very interested in discovering knowledge within the data that are usually systematically collected for a long time, sometimes tens of years [5]. Customers express their opinions which can be positive or negative, or which can be graded using a certain scale [10].

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 211–218, 2011.

One of questions can be: *What is significant for including a certain opinion into one of categories like satisfied or dissatisfied customers?* In other words, what attributes are relevant? Obviously, the answer is hidden in the words and phrases, and how they are used. The dictionary should not be too large, however, it is not easy to assess its size in advance. Human beings can more or less reliably and easily recognize the meaning of a presented opinion, however, if the opinions are written in many – more than two or three – foreign languages, the task becomes very difficult and not too many people know more languages sufficiently. Contemporary, a lot of institutions and business organizations are engaged in a trade at many countries and continents, and the problem with processing the large heterogenous data volumes is now very current.

2 Data Description

The text data used in the experiments contained opinions in many languages of several millions customers who – via the on-line Internet service – booked accommodations in many different hotels and countries. The on-line hotel reservation web has the hotels organized in a hierarchy *continent-country-city-(city district)-hotel*. Besides the information about the hotel prices, facilities, policies, terms, and conditions, the web contains user reviews related to their stay in a given hotel. The reviews cannot be entered by any person but only by the people that made a reservation through the web and stayed in the hotel. Each review consists of identification of the reviewer, his or her overall evaluation (a number on a 10-point scale) and the review text. The identification includes the type of the customer (solo traveller, family with young children, mature couple, and so like), the country and city where the reviewer comes from, and the date of the review. The review texts have two parts – a negative and positive experience with the hotel, both written in a natural language. The data was collected from more than 100 countries, 25,000 cities and districts, 108,000 hotels, and contained about 5,000,000 reviews written in many languages.

Such a big number of labeled examples enables to create sufficient training sets for an interesting number of different languages. Because of the policies of the hotel reservation web, the samples are labeled as positive and negative relatively carefully. On one hand, they are often written quite formally; on the other hand, most of them embody all deficiencies typical for texts written in natural languages (mistypings, transposed letters, missing letters, grammar errors, sometimes combinations of two languages in one item, and so on). Often, languages that normally use diacritic (for example, Czech, French, Spanish, and others) are used sometimes with and sometimes without it (written in the plain ASCII/ANSI code-page).

Labeling the reviews by the country of the reviewer enabled to automatically extract texts originated in different countries. However, belonging to a particular country doesn't necessarily mean that the reviewer used the language of that country. This means that, for instance, some reviews written by Czechs were written also in English, Slovak, or German or in combination of more languages. Just one copy of an illustrative example, Czech-English mixture, with all original mistakes: '*Co se mi nelíbylo byl hluk z ulice. There was litlle bit noise from street.*'

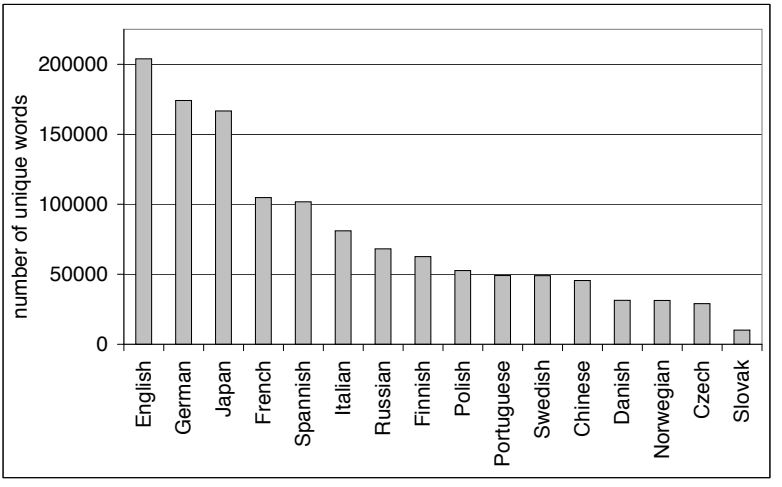

Fig. 1. Number of unique words for processed data sets

In addition, a reader of those reviews can see not only Latin alphabet: Chinese, Hebrew, Japanese, Korean, Russian, Serbian, Thai, and others. This introduces additional difficulties because each such language has its own specifics [6]. For example, Japanese texts used all three main scripts: *Kanji* (individual characters) and syllabic *Hiragana and Katakana* (in addition, they also contained sentences written in English). Another difficulty related to automatic processing of Chinese and Japanese text data is that, unlike many other languages, they do not have explicit whitespace between words. As a result, a special form of word segmentation, which is a difficult problem in these languages, is normally required before further processing [7]. Here, the authors used the same method for all languages and consequently sometimes, instead of strictly individual words, certain phrases were separated. Using a simple approach described further, the list of important words in Japanese contained, for example, 'words' such as *narrow room*, *small room*, or *since there is no elevator*.

Altogether, up to this time, there were processed the following languages: *Croatian, Czech, Danish, Finish, French, German, Hebrew, Italian, Japanese, Korean, Lithuanian, Norwegian, Polish, Portuguese, Russian, Slovak, Spanish, Swedish, Thai,* and *Chinese*. Some other data were not processed up to now because they are, for example, *Spanish*, but from *Argentina* and maybe from other South American countries. Similarly *Brazilian Portuguese*. Also, *Bassa, Catalan, Dutch, Esperanto, Estonian, Hungarian, Lithuanian, Romanian, Slovene*, and *Turkish* were not processed up to now, too, because their number of samples was too small. As a rarity, one positive review is written in *Avestan*, which is an archaic East Iranian (probably dead) language.

3 Text Document Pre-processing and Representation

The initial pre-processing of all languages was very simple and standard, based only on creating a *bag-of-words* and, consecutively, a *dictionary* from words in text documents

available [8]. The pre-processing did not use removing *stop-words* because up to now, due to many different languages and providing the same conditions to each of them, there was not enough time to create lists of such words – this is a task for the near future. The words were finally represented simply by their frequencies because other possibilities (for example, *TF*x*IDF)* did not bring any advantages as the preliminary experiments showed. The dictionary (as well as each document) was transformed into a multidimensional vector where individual word frequencies were used as coordinates within the abstract space with each dictionary word as one of dimensions. Typically, the individual vectors were very sparse because of the very large dictionary for every language and small number (in order of tens) of words in each review. For example, the collection of ca 17,000 Czech texts had about 29,000 unique words, 57,000 Russian texts 68,000 unique words, 356,000 Italian texts 81,000 unique words, 470,000 Spanish texts 102,000 unique words, and the largest collection of 1,919,000 English texts 204,000 unique words.

Table 1. The most significant words for Czech and Russian

Czech			**Russian**		
Original	%	Translation	Original	%	Translation
špatně	100	wrong	расположение	100	location
nedostatečné	100	insufficient	хороший	88	good
nemožnost	99	impossibility	отличный	81	excellent
nedostatečná	99	insufficient	приветливый	78	friendly
chyběl	99	missing	уютный	75	comfortable
chybějící	98	missing	отличное	73	excellent
koberec	98	carpet	доброжелательный	71	friendly
stěny	98	walls	отзывчивый	70	responsive
netekla	98	no water	месторасположение	69	location
zápach	97	smell	прекрасный	67	beautiful
nelíbila	97	dislike	тихое	66	quiet
nepříjemné	97	unpleasant	доброжелательность	66	generosity
nedostatečně	97	insufficiently	просторный	65	spacious
nefungoval	96	not working	природа	65	nature
nefunkční	96	not working	приятный	65	pleasant

The simply created bag-of-words suffered from obvious, commonly known shortages, for example, containing several variants of the same word – the experiments, however, had no available batch-mode stemming tools for most of all languages. Therefore, the authors decided to use all word forms because here the goal of experiments was *not to reach the best possible classification*, which is often an intention. The dimensionality of dictionaries also increases for languages that use diacritical marks (accents). Some reviewers often omitted these symbols and therefore both versions of the same word appeared.

4 Creating Dictionaries of Significant Words

For each processed language, the main goal was building a dictionary containing the words that were significant for expressing the *positive* or *negative* opinion, as well as to reveal how each word contributed to the positivity or negativity. Knowing the labels of the available samples, the authors decided to employ a decision tree generator [2] which constructs a classifier that can separate positive and negative reviews. Decision trees typically use only part of attributes and represent a set of rules. Having a set (dictionary) of a huge number of attributes (words), only a fraction of them can be decisive. One of the most popular tree generators is the algorithm that builds a tree using minimization of entropy – that is, recursively splitting the original heterogenous set of samples into homogeneous subsets. The splitting is driven by those attributes that provide the highest entropy decrease. Here, such attributes are words that direct the labeled samples to leaves as homogeneous as possible (subsets containing, if possible, only items from one class). The entropy $H(X)$ of a discrete random variable X with possible values $\{x_1, \ldots, x_n\}$ is defined by the following equation:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_b p(x_i). \tag{1}$$

If $p(x_i) = 1$ or 0, the entropy is zero, which means that a subset contains items only from one class and no items from other classes. If (for two classes) the items in a subset are divided 50% : 50%, the entropy is maximal. The most significant attribute is in the tree root, other important ones on levels close to the root. Therefore, the position of a word in the tree determines its significance. The tree asks its root each time, so this word has the 100% importance. Other words on the following tree levels may gradually be less and less important from the splitting point of view because the classification does not ask them each time, depending on a selected branch. Naturally, the classification accuracy must be as high as possible, otherwise the decisive words in the tree nodes cannot be selected reliably. In this way, the classification accuracy plays also an important role and a big number of labeled training samples is necessary. After building the tree, all significant words are known and the dictionary with words weighted by their significance can be created. Such words are then rated to be important for expressing the opinion, while the rest of them plays a negligible or no role and from the statistical viewpoint represents noise. The dictionary size (its number of unique words) is given by the tree size (its number of interior, decisive nodes). As a software tool, the authors applied the efficient c5/See5 decision tree [3].

5 Experiments and Their Results

After the initial data pre-processing, it turned out that for most of the languages the data volume was too large to be processed because of the memory requirements. The authors decided to split the primary datasets into smaller subsets using a random selection of 50,000 samples per each subset. In addition, removing words with frequency < 2 decreased the dimensionality (experiments showed that even after this filtering the results

Table 2. The most significant words for Italian and Spanish

Italian			Spanish		
Original	%	Translation	Original	%	Translation
mancanza	100	lack	no	100	not
non	97	not	demasiado	76	too
scarsa	95	poor	falta	76	absence
troppo	79	too	poco	75	bit
assenza	78	absence	olor	70	smell
carente	75	lacking	oía	70	heard
rumorosa	75	heavy	escasa	70	insufficient
odore	72	odor	oye	69	hear
poca	71	little	mala	69	bad
po	71	bit	dificultad	69	difficulty
sarebbe	70	would	excesivo	69	excessive
posizione	70	location	debería	69	I should
pò	68	bit	ruido	69	noise
rumorosità	67	noise	mal	68	bad
poco	66	little	escaso	68	insufficient

Table 3. The most significant words for German and French

German			French		
Original	%	Translation	Original	%	Translation
lage	100	location	calme	100	quiet
gutes	78	good	pas	89	not
nicht	76	no	trop	72	too
freundliche	74	friendly	bon	70	good
freundliches	73	friendly	manque	69	absence
freundlich	70	friendly	mal	67	wrong
nettes	65	nice/kind	odeur	66	smell
nette	64	nice/kind	peu	66	little
zentral	62	central	absence	66	absence
gute	61	good	mauvaise	61	bad
schönes	61	nice	insuffisante	60	insufficient
freundlichkeit	59	friendliness	bruyant	60	noisy
zuvorkommend	56	helpful	plastique	60	plastic
tolles	56	great	excessif	59	excessive
ruhig	55	quiet	bruit	59	noise

were quite identical). Then, the decision tree was applied to all individual (sub)sets of data and the trained trees were used for creating the dictionaries of words significant for expressing the opinions. Which words belong to the positive or negative class, it is given by a relevant branch which terminates in a 'positive' or 'negative' leaf. The words were represented by their frequencies. The decision tree mostly asked if the frequency was > 0 or $= 0$, which was, in fact, the binary representation. However, sometimes the

Table 4. The most significant words for Japanese and English

Japanese			English	
Original	%	Translation	Original	%
部屋が狭く	100	narrow room	location	100
部屋が狭い	100	the small room	friendly	80
エレベーターがないので	100	since there is no elevator	not	77
強いて言えば	100	if I'm forced to say	excellent	73
結局	100	eventually	helpful	67
残念	99	shame	spacious	62
残念でした	99	too bad	friendliness	59
清潔	99	clean	beautiful	57
駅から近く	99	near station	comfortable	55
便利でした	99	was useful	nice	55
スタッフも親切でした	99	the staff was kind	conveniently	54
部屋は清潔	99	rooms are clean	convenient	53
狭い	98	narrow	fantastic	53
快適に過ごせました	98	we had a comfortable	good	52
ロケーション	98	location	proximity	52

real frequency value played its role, therefore the results were slightly worse for the binary encoding.

Each of the 50,000-samples subsets gave almost the same list of words; usually, only the position of the words fluctuated. The number of selected significant words was only around 200. In the tables Tab. 1 – Tab. 4, because of the limited article size, only the first 15 words with their average position weight (column %) is shown for selected languages. As for reliability of this selection, the accuracy of the classifier was typically between 85-93% – lower values for lower numbers of training samples depending on a language. The accuracy estimate of the classifier performance was given by the 5-times 10-fold-crosvalidation procedure. To avoid the classifier overfitting, the c5/See5 default global pruning confidence factor 25% was applied.

6 Conclusions

This research describes a method of extracting significant words from unstructured textual customer reviews written in various natural languages. The word significance was given by expressed opinions of provided services (hotel accommodations), positive and negative. Because of the very large data volumes in many different languages, the suggested approach chose a relatively simple but unified way to pre-process the data. The application of the c5/See5 decision tree generator selected the significant words according to their belonging to the positive or negative opinion (the classification itself was not the goal). At the same time, the position of the tree nodes (that contained the words) represented the significance degree of the individual extracted decisive words – the most important one was in the root. This approach noticeably decreased the number

of all words (in the order of 10^4 to 10^5) to about 200. Such an approach can be used for building dictionaries that include only relevant information and are not too large. The dictionaries then enable more effective analysis of the data, which is attractive both for commercial and non-commercial entities. Selecting significant words, as described above, was succesfully applied to building dictionaries with words expressing opinions in natural languages [4]. In addition, those dictionaries were further used for looking for significant phrases (submitted to another conference). The future research is going to focus on further improvement of the suggested approach, mainly on the data preprocessing phase, which is complicated by the large language variety and high data volumes. Another part will include more detailed opinion analysis when the reviews can be graded, having more than two classes.

Acknowledgements. The research work published in this paper was supported by the Research program of Czech Ministry of Education VZ MSM 6215648904. The authors thank Eva Dařenová and Eva Miškovičová for their valuable help with analysing text documents written in several foreign languages.

References

1. Berry, M.W., Kogan, J. (eds.): Text Mining: Applications and Theory. John Wiley & Sons, Chichester (2010)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2007)
3. c5/See5 (2011), `http://www.rulequest.com/see5-info.html`
4. Dařena, F., Žižka, J.: Text Mining-Based Formation of Dictionaries Expressing Opinions in Natural Languages. In: Proceedings of the 17th International Conference on Soft Computing Mendel 2011, Brno, June 15-17, pp. 374–381 (2011) ISSN: 1803-3814
5. Liu, B.: Web data mining: Exploring Hyperlinks, Contents, and Usage Data. In: Opinion Mining. Springer, Heidelberg (2006)
6. Nie, J.Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies 3(1), 1–125 (2010)
7. Peng, F., Huang, X.: Machine learning for Asian language text classiffication. Journal of Documentation 63(3), 378–397 (2007)
8. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 1, 1–47 (2002)
9. Shmueli, G., Patel, N.R., Bruce, P.C.: Data Mining for Business Intelligence. John Wiley & Sons, Chichester (2010)
10. Žižka, J., Dařena, F.: Automatic Sentiment Analysis Using the Textual Pattern Content Similarity in Natural Language. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 224–231. Springer, Heidelberg (2010)